Efficient multiple testing for large structured problems

Meijer, R.J.

1 Introduction

1.1 Introduction

Modern day technology allows us to collect and store more data than ever before. Although the abundance of data offers great opportunities with respect to formulating and testing all kinds of hypotheses and with respect to discovering and modeling patterns that can be used for future prediction, these opportunities do not come without risks. If the data is explored in a too naive manner, there is a very large probability that, at least part of, the actual findings will not be reproducible in future research. A risk already pointed out by Ioannidis (2005) in his famous article entitled “Why most published research findings are false”. The reason for such chance findings mainly lies in the nature of statistics, which is sometimes referred to as “the science of uncertainty”. For every supposed finding, there is a possibility that this finding only occurred by chance in the sample under study, while it is not present in the population at large. When the number of hypotheses or the number of parameters within a model increases, the possibility of actually ending up in such a situation increases as well.

Profiting from all available data, while preventing chance findings from occurring, will call for advanced statistical techniques. The development of such techniques will be the subject of this thesis. We focus both on multiple hypothesis testing and on the construction of prediction models. Throughout the thesis, emphasis is placed on the development of methods that are efficient, both from a power perspective and from the perspective of computational feasibility.

Although the use of our proposed procedures is not limited to a specific data type, our procedures will be demonstrated on high-dimensional biological data and for that reason this type of data will also serve as an example throughout this introduction. Henceforth, let us assume that our data consists of p gene expression measurements for n individuals for which we also have some phenotypic information, such as whether the individual suffers from a certain disease. These gene expression measurements tell us at what rate several genes, which are just specific regions in our DNA, are transcribed into RNA copies, which can subsequently be translated into proteins (or other functional gene products). Because these proteins can have a major impact on many processes in our body, it seems reasonable to hypothesize that (some of) the genes can have an effect on our phenotypic outcome variable.

In the coming sections the principles of hypothesis testing, and in particular multiple testing procedures, and statistical modeling will be discussed in more detail, motivated from the data setting described earlier. Additionally, one section will be used to address a different challenge that arises when the size of the data sets increases: the challenge of computational feasibility. The last section provides an outline of the thesis as a whole.

1.2 Hypothesis Testing

In hypothesis testing, the aim is to decide whether or not some pre-specified hypothesis, called the null-hypothesis, can be rejected based on evidence that is present in the data. Null-hypotheses come in many forms, but in this introduction we will look at a specific null-hypothesis that will often serve as an example throughout this thesis; the so-called self-contained null-hypothesis (Goeman and Buehlmann, 2007) as found in gene set testing. This hypothesis states that, given a gene set G, consisting of one or more genes, none of the genes within G are associated with the outcome variable. If this hypothesis can be rejected, this thus tells us that there is at least one gene within this group that is in fact associated with the outcome, where this outcome could for example be disease status.

Deciding which gene set or sets are interesting to test will often be done based on prior knowledge. Sometimes, the interest will be in discovering single genes that are important in the development of a specific disease, whereas in other cases the focus will be on larger gene sets. For many genes it is known that they belong to the same cellular component, are involved in the same biological process, or have the same molecular function. Such information can be looked up in annotation databases, such as the Gene Ontology database (Ashburner et al., 2000). In this way, biologically interesting gene sets can be formed. Often more than one gene or more than one gene set will be of interest, which results in the formulation of several null-hypotheses.

1.2.1 Type of test

In order to decide whether a null-hypothesis can be rejected, a statistical test is needed. Which test can best be used will usually depend both on the exact form of the null-hypothesis and on the expected alternative hypothesis. Even though the alternative hypothesis does not have to be specified explicitly, different tests can have different power against different alternatives.

Let us first formulate our self-contained null-hypothesis more precisely as:

$$H_0\colon \beta_i = 0, \text{ for all } i \in G \subseteq \{1,\ldots,p\} \tag{1.1}$$

where each β_i is a regression coefficient corresponding to gene measurement x_i, with $i \in G$, in a regression model that has y as its outcome variable and all genes within G as its covariates. When the outcome is disease status, this could for example be a logistic model, but different models are possible for different outcome variables. When null-hypotheses of this form are tested, usually the alternative hypothesis is not specified and it is assumed that the true model is best estimated by the model that maximizes the corresponding likelihood. The three most well-known tests, which are the Wald test, the likelihood-ratio test and the score test, are all based on the principle of maximum likelihood estimation. As long as the number of covariates $|G|$, where $|G|$ denotes the size of group G, in our model is sufficiently smaller than the number of observations n, all these tests can be used and will give approximately the same results.

However, as soon as the number of covariates comes close to or even exceeds the number of observations, maximizing the likelihood is no longer straightforward, because the resulting regression model will be unstable or cannot be fitted at all. In that case, tests that are developed for such high-dimensional situations can be used, such as the global tests developed by Goeman et al. (2004) or Mansmann and Meister (2005). These tests will have optimal power against the alternative in which many of the genes are (weakly) associated with the response. When fewer, larger effects are expected, other tests, such as for example a Simes' test (Simes, 1986) might be more powerful. Alternatively, one can decide to test smaller groups of genes or even single genes, because testing large gene sets might not be advantageous in such situations.

In each statistical test, a test statistic T is defined for which the distribution under the null-hypothesis can be derived either exactly or approximately. Subsequently it can be checked whether the value of the test statistic for the observed data, T_{obs}, is probable given this null-distribution. This can be done by calculating the corresponding p-value, which is the probability of observing a result at least as extreme as T_{obs}, given that the null-hypothesis is true. The p-value can thus be seen as an informal measure of evidence against the null-hypothesis; the smaller the p-value, the less likely it is that the null-hypothesis is true. However, it is important to note that a small p-value can occur simply by chance even if a true null-hypothesis is tested.

The decision whether or not to reject the null-hypothesis is now made by comparing the corresponding p-value to a pre-specified significance level α, which is often set to 0.05. If the p-value is smaller than the chosen α-level it is decided that there is enough evidence to reject the null-hypothesis. Otherwise, the null-hypothesis is accepted. Two types of errors can be made here, namely a type I and a type II error. A type I error occurs when the null-hypothesis is wrongly rejected. This happens with probability at most α since the p-values are approximately uniformly distributed under the null-hypothesis.

A type II error is made when the null-hypothesis is accepted while it should have been rejected. This can for example happen when the sample size or the magnitude of the true effect is not large enough or when the test has too little power against the actual alternative. Although both types of errors should ideally be prevented, type I errors are usually considered to be the most problematic errors, because a highly significant result appears definitive and has the effect of stopping further investigation, as was already pointed out by Bakan (1966). Whereas a type I error will only be made with probability at most α when one hypothesis is tested on a significance level of α, this changes as soon as more than one hypothesis is tested.

1.2.2 Multiple Testing

In most situations, researchers are not interested in only testing whether one specific gene or one specific gene set is related to a given biological phenomenon, but they want to investigate the influence of many genes and/or gene sets. Suppose that instead of one null-hypothesis we have a set of m null-hypotheses $H_1, \ldots, H_m$ that we would like to test. Given these m hypotheses, we know that there is an unknown number m_0 of true null-hypotheses among them while the remaining $m_1 = m - m_0$ are the hypotheses that we wish to reject. Suppose that for each hypothesis H_i we obtained a p-value p_i by applying an appropriate test. If each of these p-values would be compared to the same α-level of, say, 0.05, the expected number of type I errors would equal $m_0 \times \alpha$, which can clearly become very large when the number of true hypotheses is large as well. To prevent type I errors, or to at least limit their number, we need some multiple testing procedure.

Most multiple testing procedures can be subdivided into two categories; those procedures that control the *familywise error rate (FWER)* and those procedures that control the *false discovery rate (FDR)*. If the FWER is controlled on level α, this means that the probability of making *any* type I error is bounded by α. If the FDR is controlled at level α, some type I errors will be allowed as long as the expected proportion of type I errors among all rejections will stay below α. In general, controlling the FDR will result in more rejections than controlling the FWER. This is especially the case when many hypotheses are tested and if a substantial part of these hypotheses is false. Although FDR controlling methods are for this reason rather popular, especially in a genomic context, in this thesis we will mainly focus on FWER control. Not only because some settings, for example validation experiments, ask for strict control of the number of type I errors, but also because the property that, with probability at least $1 - \alpha$, no type I errors have been made, allows for certain reasoning that does not apply to FDR control.

There exist several different FWER controlling methods. As with choosing an appropriate test, choosing an appropriate FWER controlling procedure also depends on the hypotheses under study. If many non-logically related hypotheses are tested, such as hypotheses of the form

$$H_0\colon \beta_i = 0 \tag{1.2}$$

that state there is no association between one individual gene i and the outcome variable, standard methods such as the methods of Bonferroni (1935), Holm (1979), Hochberg (1988) or Hommel (1988) can be used. These methods test all hypotheses simultaneously, but on smaller individual α-levels than the α on which the FWER should be controlled. FDR controlling methods that work similarly are the methods of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001).

As soon as logically related hypotheses are tested, other multiple testing procedures can be more powerful. Logically related hypotheses arise for example when several nested gene sets are tested. This will often happen because both general biological processes, relating to large gene sets, and specific processes, relating to smaller gene sets or even single genes, can be of interest. If a single gene is found to be associated with the outcome, the same must hold for all gene sets of which it is part, and similarly, if a gene set is found to be associated with the outcome, this has to be because the same must hold for at least one of the genes within this gene set. This leads to so-called *restricted combinations* among the hypotheses, as first discussed by Shaffer (1986), which means that not all combinations of true and false hypotheses can exist, but the falsehood of one hypothesis will often imply the falsehood of others.

When logical relations exist between the hypotheses, a multiple testing procedure can take advantage of these by recognizing that not all remaining hypotheses can be simultaneously true and adjusting the individual α-levels accordingly, which leads to more powerful procedures. Besides, the logical relations can be used to structure the testing problem. If nested gene sets are tested, one can for example use a procedure that starts by testing the larger sets, and which only continues testing the smaller sets once there is evidence that these could prove to be significant. This is done in hierarchical testing procedures, such as procedures developed by Goeman and Mansmann (2008) or Meinshausen (2008). Because in each step of the procedure only a subset of the original hypotheses is tested, all individual tests can be done on larger α-levels which increases the procedure's power.

To be able to use logical relationships that exist between hypotheses or to carry out the tests in a particular order while still controlling some error criterium, it will usually not be enough to know that most of the current rejections are true rejections, but one needs that, with high probability, all current rejections are true rejections, as will be explained in chapter 5 of this thesis. For that reason, this thesis focuses on FWER controlling multiple testing procedures instead of on FDR controlling procedures. The focus will be on developing novel multiple testing methods for several types of hypotheses. The aim is to make full use of the structure and underlying logical relations between the hypotheses.

1.3 Prediction modeling

In statistical modeling, the aim is to develop a model that describes the relation between a number of covariates and an outcome variable. Such models can be used to gain insight in the phenomenon under study, but very often they are also used to predict future outcomes. Various outcomes can come to mind, such as the underlying diagnosis, the expected response to treatment or the expected survival time. A model can for example be developed to see how gene expression levels, measured at time of entry in a study, influence the survival time for patients that are diagnosed with a certain illness and are kept under surveillance for a long period of time. For new patients, for whom only the gene expression levels at baseline are known, this model can now be used to give an estimate of for instance the 5-years survival probability.

Although the usefulness of prediction models is apparent, developing a valid prediction model can be difficult. Especially when the number of covariates is large, there is a considerable risk of constructing a model that predicts very well on the current data set, but that has little or even no predictive power for similar new data sets. This phenomenon is called overfitting. A link can be made here with multiple testing problems, where we already saw that a naive approach will usually result in many finding that will not be reproducible in new experiments. To try to prevent overfitting, advanced modeling techniques can be used, such as we will describe in the next subsection. Furthermore, a measure of predictive performance should be used to predict how reliable the model will be for new data sets. To prevent chance findings, it is very important that this measure indeed measures the expected performance on a *new* data set and not on the existing data set.

1.3.1 Type of model

Before a model can be properly constructed and tested, one needs to decide what type of model will be useful. Several factors play a role in this decision. Firstly, the type of outcome variable is important, since different models are needed for dichotomous, continuous or for example survival outcomes. Secondly the number of covariates will be of influence, since models in which the number of covariates exceeds the number of observations ask for different modeling techniques than lower dimensional models. The expected effects are also important; will these effects be linear, or is a deviation from linearity expected which can be modeled for example by splines? An other important question is whether the objective is to identify the “true underlying model” or whether the model will mostly be used for prediction purposes, while the exact interpretability is of less importance. Ideally, the type of model one wants to fit is already decided before exploring the actual data, to make sure that this decision is made based on reasoning and not based on accidental findings in the current data set, but this strategy will in practice often be too inflexible.

When the preferred model has not too many covariates, it is common practice to fit the model by using maximum likelihood optimization as already discussed in the previous section. The same approach cannot be taken for models with many covariates however, which would for example be the case when all measured gene expression levels, which can easily range in the tens of thousands, are used as covariates. As soon as the number of covariates p comes close to the number of observations n, the resulting maximum likelihood estimates will become very unstable, and as soon as p equals or exceeds n, infinitely many solutions exist that will result in perfect prediction on the current data set, but it is clear that such models will have no predictive power at all for future data.

To prevent overfitting, penalization techniques have been developed. When such techniques are used, before maximizing the likelihood a penalty term is added to this likelihood which prevents the regression coefficients from attaining large and unrealistic values. The penalty term can be chosen in different ways, resulting in regression models with different behavior. Two well-known penalized regression techniques are lasso regression (Tibshirani, 1996) and ridge regression (Hoerl and Kennard, 1970). Whereas lasso regression performs variable selection, which means that only a subset of the original covariates will be present in the final model, ridge regression leaves all covariates in the model but all regression coefficients will be very small. Lasso models are often used when one wants to obtain parsimonious and preferably interpretable prediction rules. Ridge models will more often be used when prediction is the sole focus of the model.

Once the model is fitted, it should be checked whether it is a useful model, in other words whether we expect it to be a reproducible model, or whether it only describes the current data set. For this purpose, measures of predictive performance have been developed.

1.3.2 Predictive performance

There are various performance measures that can be used to assess the performance of a statistical prediction model. Some measures directly compare the outcomes predicted by the model to the actual outcomes. The differences between these values indicate the goodness-of-fit of the model, where smaller distances between the predicted and actual outcomes indicate a better model. But other, more indirect measures are possible as well. In case of binary outcomes, one can for example check for all patient pairs of which one patient experienced the event (the case) and one did not (the non-case) whether the model indeed predicted a higher probability of experiencing the event for the case relative to the non-case. The more correctly predicted pairs, the better the model, which is the reasoning underlying the often used receiver operating characteristic (ROC) curve. Steyerberg et al. (2010) provide an overview of different performance measures that are frequently used in medicine.

When the value of a performance measure is calculated using the same data as the data used in fitting the model, this value only indicates how well the model predicts for this particular set, but it cannot be used to assess how well the model will perform on new data, since a high value could easily be the result of overfitting. To transform it into a measure of *predictive* performance, the value should be calculated on data that was not used in fitting the model. If a similar data set is available, this data can be used to assess the method's predictive performance. However, usually such an independent data set will not be available. Another idea is to split the current data set in a train and a test part. The model is fitted on the training set and the performance measure is subsequently calculated by means of the test set. A systematic way of splitting the data several times into a train and test set is by using cross-validation. When cross-validation is used, the data is divided into disjoint subsets of (approximately) the same size. Each time, one of the subsets is used as test set, while the remaining subsets are used as training set. The model is fitted on the training set, evaluated on the test set, and this is done repeatedly such that each subset once serves as test set.

When cross-validation is used, the resulting value for the performance measure will no longer be too optimistic, and it can be seen as a true measure of predictive performance. It is thus an important tool in the prevention of chance findings. One disadvantage of cross-validation is however that, instead of fitting one model, many models have to be fitted, which can be very time-consuming.

1.4 Computational feasibility

When the number of hypotheses and the complexity of regression models increases, preventing the occurrence of chance findings becomes not only more challenging from a methodological point of view, but also from a computational point of view. This computational challenge is not limited to the prevention of chance findings but is encountered more generally since statistical methods are expected to be applicable in situations with ever increasing amounts of data. This expectation will not always be reasonable however.

Once a situation arises in which problems regarding computation time are present or are to be expected, the first step is to gain insight in the magnitude of the problem. It might be that the current method will have to be updated or maybe even has to be replaced completely, but this will only become clear after careful examination of the problem at hand. An important first question is how fast the computation time will increase when the size of the input increases. When the computation time will increase linearly with the size of the input, which would mean for example that if twice as many hypotheses are tested, the process take twice as long, there is much less to worry about than when the computation time increases cubically or even at a non-polynomial rate. In computer science, the response to changes in input size is often denoted by using big O notation. If an algorithm is of quadratic order for example, which is denoted by $O(n^2)$, this means that the growth rate of the computation time with respect to the input size n will be bounded by a second order polynomial.

If the order of the problem is not too large, say, smaller than quadratic, there are often possibilities to keep the computation time within reasonable bounds. Firstly, most programs can be written more efficiently, for example by avoiding unnecessary copying of data or by applying vectorization if possible. There is also the possibility to write (part of) the code in a lower level programming language, which will usually result in a faster program. Furthermore, a lot of free and commercial software is available that is highly optimized in terms of computational performance. Checking whether an efficient solution for the problem at hand has already been found can save a lot of time. If these small changes are not yet sufficient, there might be the possibility of using parallelization techniques. If the problem can be split up into smaller, independent, subproblems, this will often be an option.

Although the possibilities just described can have a substantial impact on the computation time, it will usually not be enough when the algorithm has a large growth rate. If the order of the desired method is large, one can first look whether this order can be brought down by using a different calculation technique. A different mathematical approach can sometimes lead to the desired result. In other situations, asking advice from experts, such as computer scientists, can be beneficial. Often such a reduction will not be possible however, or the new order will still be too large. In that case, one can still consider whether the results can be approximated rather than exactly calculated, which can often save a lot of time. When this is not possible or not desirable, there is always the option to decide that the original plan is impossible. This can either result in the decision to accept that one has stumbled upon an unsolvable task, or the decision to take a totally different approach. Or one can decide to look for middle ground by solving the problem in the way that was anticipated, while additionally requesting restrictions on the input size. Even though the requirement for restrictions on the input size might sound unsatisfactory, this can sometimes even lead to better stated problems, if it results in considering only those hypotheses or only those variables that are truly of interest.

Each computational problem thus asks for its own solution. Throughout this thesis many computational challenges will be encountered, each of which will be overcome by using one or more of the techniques mentioned above.

1.5 Outline of the thesis

This thesis is a collection of five articles and one book chapter. In principle, all coming chapters can be read in any preferred order, since all documents are meant to be self-contained. However, we feel that the current ordering does most justice to the strong connections that exist between all chapters. Chapter 2 until 5 will solely focus on multiple testing methods. Chapter 6, which is the original book chapter, is a chapter in which no new methodology is introduced, but which brings together the fields of multiple testing and statistical modeling. In chapter 7, the focus is solely on model building and model evaluation.

In chapter 2, we describe an algorithm to make Hommel's FWER controlling procedure more efficient. By using this algorithm, finding adjusted p-values for n elementary hypotheses, which can be hypotheses of the form given in equation (1.2), can be done in $O(n\log(n))$ steps instead of the $\Theta(n^2)$ steps that are needed in Hommel's original procedure. This makes the procedure feasible for a very large number of hypotheses. Whereas Hommel's procedure focuses solely on the elementary hypotheses, we also show how one can determine whether an intersection hypothesis, which can be of the from given in equation (1.1), can be rejected by the closed testing procedure (Marcus et al., 1976) in combination with a Simes' test and we develop an algorithm to calculate confidence sets for the number of true or false hypotheses within any arbitrarily chosen set of elementary hypotheses, following the example of Goeman and Solari (2011).

In chapter 3, a novel multiple testing procedure is introduced for the situation in which the elementary hypotheses are ordered in either space or time. Given such an ordering, specific intersection hypotheses become hypotheses of interest, namely those intersections that consist of consecutive elementary hypotheses. We call these hypotheses the region hypotheses. Because all regions, of different lengths and in different positions, are potentially interesting, we propose a method that tests all possible region hypotheses (including all elementary hypotheses) while controlling the FWER. The procedure is a step-down procedure, meaning that the hypothesis corresponding to the largest possible region, i.e. the intersection hypothesis of all elementary hypotheses, is tested first and if this hypothesis can be rejected, we continue with further specifying the exact location/locations of the effect present. The procedure uses the logical relations that exist between the different region hypotheses to gain power and is based on the sequential rejection principle of Goeman and Solari (2010). As in the previous chapter, again an algorithm is provided to calculate confidence sets for the number of true or false hypotheses within every arbitrarily chosen set of elementary hypotheses.

The multiple testing procedure that is proposed in chapter 4 can be seen as a more general variant of the region procedure. This new procedure can be used for testing hypotheses that are structured in a directed acyclic graph (DAG). The set of region hypotheses could be an example of such a graph structure, but because all DAG structures are possible, the DAG method is much more flexible than the region method. An interesting example of a graph structure that can be tested with this new method is the Gene Ontology graph. As before, a top-down approach is used and confidence sets for the number of true or false hypotheses within any set of elementary hypotheses can be constructed.

Whereas the region and the DAG method are top-down procedures, the multiple testing method that is introduced in chapter 5 is a simultaneous (i.e. non-hierarchical) method, which means that no particular testing order is used but all hypotheses can be tested simultaneously in every step. This method can again be used on any DAG structure and can be seen as a modified version of the well-known method by Holm (Holm, 1979) in which the logical relations between the hypotheses are used to gain power. Apart from introducing a new FWER controlling procedure, chapter 5 also discusses the differences between FWER and FDR controlling procedures in the context of gene set testing and suggests a specific way to interpret and summarize results from gene set testing procedures.

Chapter 6 revolves around the research question “How to select important covariates from a large set of candidates?” and through this question, the connection between multiple testing and model building is addressed. We give an overview of well-known multiple testing and variable selection procedures and discuss why variables selected for a (multivariate) prediction model and variables selected by a (univariate) multiple testing procedure can be quite different. Furthermore, we discuss what it means in practice if a variable selection procedure has the “oracle property”, i.e. the property that it will only select those variables that are present in the underlying true model.

In the final chapter, we describe an approximation method for cross-validation, which is a resampling method that is frequently used in both model building and in evaluating the predictive performance of the final model. Whereas true cross-validation requires fitting a prediction model multiple times, each time on a slightly different data set, our approximation model uses a Taylor expansion around the estimate of the full model in order to approximate the cross-validated estimates. In this way, these estimates can be obtained without refitting the model, which makes this method much more efficient than actual cross-validation. The method can be used for generalized linear models and Cox' proportional hazards model with a ridge penalty term.

Throughout the thesis, high emphasis is put on the computational efficiency of our methods. To facilitate the actual use of all procedures, accompanying software for all described procedures can be found in either the `R`-package `cherry` or the `R`-package `penalized` which can be freely downloaded from the `CRAN` repository.